*Logistic Regression*

* In simple terms, “What’s the probability that something belongs to a particular class?”
* For example:
* Will the email be Spam or Not Spam?
* Willa customer churn or not churn?
* Will the car sell or not sell?
* It predicts the probability between 0 and 1, then classifies:
* If it’s >= 0.5, classify as 1 (Positive Class).
* If it’s < 0.5, classify as 0 (Negative Class).
* Logistic Regression squeezes the linear regression output into the range between 0 and 1 using this magic curve called the sigmoid.
* \sigma(z) = \frac{1}{1 + e^{-z}}
* This function takes any value and transforms it to a smooth probability.
* It’s historical. It’s actually used for classification, but the underlying math fits within a regression framework.
* Logistic Regression is a statistical and machine learning technique used for classification problems, especially when the output (target is binary) – like yes/no, true/false.
* It can also be used for multi-class classification, but its foundation is in binary classification.

How It Works?

* Takes your features.
* Computes a weighted sum (similar to linear regression).
* Applies a transformation (sigmoid curve) to squash that output into a range of 0 to 1.
* Interprets that number as a probability.
* It doesn’t say just ‘Yes’ or ‘No’, it tells us how confident it is.

Pros

* It’s simple and fast.
* Outputs probabilities, not just hard classifications.
* Works well when classes are linearly separable.
* It’s interpretable – you can examine the impact of each feature.

Used In

* Spam detection
* Customer Churn Prediction
* Credit Risk Modeling
* Disease Diagnosis
* Marketing

ROC Curve & AUC Curve

* ROC stands for Receiver Operating Characteristic.
* It is a graphical plot used to evaluate the performance of a binary classification model at various threshold settings.
* It plots:
* True Positive Rate (TPR) = Sensitivity / Recall
* TPR = TP/TP + FN
* False Positive Rate (FPR)
* FPR = FP/FP + TN
* AUC = Area Under the Curve (ROC Curve)
* Measures the entire 2D area underneath the ROC curve.
* AUC ranges from 0 to 1:
* 1.0: Perfect Classifier
* 0.5: No better than random guessing
* < 0.5: Worse than guessing (usually a bug)
* It helps us compare classifiers, especially when:
* The dataset is imbalanced.
* We’re looking for model performance independent of threshold.
* We care about the ranking of predictions, not just the class labels.

|  |  |
| --- | --- |
| AUC Score | Model Performance |
| >0.9 | Excellent |
| 0.8-0.9 | Good |
| 0.7-0.8 | Fair |
| 0.6-0.7 | Poor |
| <= 0.5 | Useless (random or worse) |

* When to use ROC-AUC:
* We don’t want to rely on Accuracy.
* Data is imbalanced.
* We want to compare multiple models.
* We care about both true positives and false positives rates.